多任务学习的不平衡 SVM+算法 *

周国华[1,2]，过林吉[1]，殷新春[2]

(1. 常州轻工职业技术学院 信息工程与技术学院，江苏 常州 213164；2. 扬州大学 信息工程学院，江苏 扬州 225127)

摘　要：处理不平衡数据分类时，传统支持向量机技术(SVM)对少数类样本识别率较低。鉴于 SVM+技术能利用样本间隐藏信息的启发，提出了多任务学习的不平衡 SVM+算法（MTL-IC-SVM+）。MTL-IC-SVM+基于 SVM+将不平衡数据的分类表示为一个多任务的学习问题，并从纠正分类面的偏移出发，分别赋予多数类和少数类样本不同的错分惩罚因子，且设置少数类样本到分类面的距离大于多数类样本到分类面的距离。UCI 数据集上的实验结果表明，MTL-IC-SVM+在不平衡数据分类问题上具有较高的分类精度。

关键词：不平衡数据；支持向量机；SVM+；多任务学习；分类

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2018.03.0276

Multi-task learning of SVM+ for imbalanced classification

Zhou Guohua[1, 2], Guo Linji[1], Yin Xinchun[2]

(1. *School of Information Engineering & Technology*, *Changzhou Institute of Light Industry Technology*, *Changzhou Jiangsu* 213164, *China*; 2. *College of Information Engineering*, *Yangzhou University*, *Yangzhou Jiangsu* 225127, *China*)

Abstract: When learning from imbalanced datasets, the traditional support vector machines (SVMs) had a low rate of identification on the minority class. Inspired by that SVM+ can utilize the additional information hidden in the training data and multi-task learning can improve the generalization performance by training multiple related tasks simultaneously, this paper proposed a new support vector machine called multi-task learning SVM+ for imbalanced classification (MTL-IC-SVM+) . MTL-IC-SVM+ incorporated the multi-task learning framework into SVM+ to hand the problem of class imbalance by applying the different penalty factors to the data, especially, the margin between the hypersphere and the minority class was as large as possible. Experiments conducted on several UCI datasets show that the proposed methods lead to very encouraging results on imbalanced datasets.

Key words: imbalanced datasets; support vector machine; SVM+; multi-task learning; classification

0　引言

支持向量机（SVM）同时以结构风险和经验风险最小化为原则，能利用核技术处理非线性识别问题，与其他机器学习方法相比，SVM 具有良好的泛化性能。但常规的 SVM 都只适应于数据平衡的分类场景，而在不平衡数据下，SVM 为达到整体数据分类误差的最小化倾向于追求多数类样本的高识别率，此时分类面向少数类样本偏移造成少数类样本的高误判率[1,2]。但在实际应用中，不平衡数据广泛存在与各个领域，如网络入侵检测、图像识别、信息检索与过滤、医疗诊断、工业过程检测等[3~5]。因此，研究 SVM 在不平衡数据分类上的应用是有必要的和值得关注的。目前，SVM 中处理不平衡数据的策略可分成基于数据采样和基于算法调整的两种。前者的代表有过采样和欠采样算法[6]；后者的代表有代价敏感学习[7]、Boosting 技术[8]和不平衡集成学习[9]等。但过采样易出现过拟合现象；欠采样易导致数据信息的不完整，同时真实的错分代价在代价敏感学习中常难以准确估计；Boosting 技术与 SVM 相结合往往伴有大的计算量；不平衡集成学习一般通过迭代的方式优化训练数据集而无法保证分类结果的全局最优解[10,11]。

近期研究表明，多任务学习通过多个相关任务的共同学习能明显提高单个任务学习的性能。同时多任务学习能有效利用任务相关性，因而对样本较少的分类情况是非常有效的[12, 13]。受此启发，本文提出不平衡 SVM+分类算法（multi-task learning based on SVM+ for imbalanced classification, MTL-IC-SVM+）。Vapnik 提出的 SVM+[14]建立在传统 SVM 模型上，但将松弛变量用修正函数的形式表示，用以挖掘样本间隐藏的结构信息。

收稿日期：2018-03-28；**修回日期：**2018-06-04　　**基金项目：**国家自然科学基金资助项目（61472343）

作者简介：周国华（1977-），男，江苏东台人，讲师，硕士，主要研究方向为智能学习、模式识别（tiddyddd@sina.com.cn）；过林吉（1982-），女，江苏武进人，讲师，硕士，主要研究方向为智能识别；殷新春（1962-），男，江苏姜堰人，教授，博导，主要研究方向为人工智能、密码学.

鉴于 SVM+算法在单任务学习中的高泛化性能，本文在 SVM+模型的基础上分别赋予多数类和少数类样本不同的错分惩罚因子，且基于分类面“大间隔”的策略，设置少数类样本到分类面的距离大于多数类样本到分类面的距离；同时参照多任务学习的框架将不平衡数据的分类表示为一个多任务的学习问题，利用相关任务间的有效信息来提高学习所得模型的泛化能力。

1　SVM+算法

为了提高 SVM 性能和减少训练所需的样本数，SVM+算法将样本的结构信息引入到 SVM 模型。与常规 SVM 中松弛变量为一实数不同，SVM+中松弛变量表示为一组修正函数。设给定样本集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_N\}$ 和其对应的类别标签 $\boldsymbol{Y}=\{y_1,y_2,...,y_N\}(y_i\in\{-1,+1\},i=1,2,...,N)$，依据属性特征的覆盖范围将训练样本划分成 t 组，每组样本及其标签可以表示为

$$\boldsymbol{D}_r=\{\{\boldsymbol{X}_r,\boldsymbol{Y}_r\},r=1,...,t\}=\{(\boldsymbol{x}_{r_i},y_{r_i})\subseteq\boldsymbol{R}^N,i\in T_r\}$$

其中:T_r 表示分组编号。SVM+使用核技术将训练样本映射到两个不同的 Hilbert 空间：（1）使用核函数 $\phi(\boldsymbol{x}_i)$ 将全部训练样本映射至决策空间 $\boldsymbol{Z}$，并对应得到决策函数（$(\boldsymbol{w}, b)$为决策函数参数））（2）使用核函数 $\phi_r(\boldsymbol{x}_i)$ 将训练样本映射至修正空间 $\boldsymbol{Z}_r$，并由此得到 r 组修正函数（$(\boldsymbol{w}_r, d_r)$为修正函数参数）），即

$$\xi_r(\boldsymbol{x}_i)=(\boldsymbol{w}_r\cdot\phi_r(\boldsymbol{x}_i))+d_r$$
$$\phi_r(\boldsymbol{x}_i)\in\boldsymbol{Z}_r,i\in T_r,r=1,...,t \tag{1}$$

SVM+中所有样本使用同一核函数映射至同一决策空间；但不同组别样本映射到修正空间时可以使用不同的核函数映射至不同的核空间。SVM+目标函数可以表示为[15]

$$\min_{\substack{\boldsymbol{w},\boldsymbol{w}_1,...,\boldsymbol{w}_r,\\ b,d_1,...,d_r}}\frac{1}{2}(\boldsymbol{w}\cdot\boldsymbol{w})+\frac{r}{2}\sum_{r=1}^{t}(\boldsymbol{w}_r\cdot\boldsymbol{w}_r)+C\sum_{r=1}^{t}\sum_{i\in T_r}\xi_i^r$$
$$\text{s.t. } y_i(\boldsymbol{w}\cdot\phi(\boldsymbol{x}_i)+b)\ge1-\xi_i^r,i\in T_r,r=1,...,t$$
$$\xi_i^r=(\boldsymbol{w}_r\cdot\phi_r(\boldsymbol{x}_i))+d_r,i\in T_r,r=1,...,t$$
$$\xi_i^r\ge0,i\in T_r,r=1,...,t \tag{2}$$

引入非负的 Lagrange 因子 $\boldsymbol{\alpha},\boldsymbol{\beta}$，SVM+的对偶问题可表示为如下二次规划问题：

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}}\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j\phi(\boldsymbol{x}_i)\cdot\phi(\boldsymbol{x}_j)+$$
$$\frac{1}{2\gamma}\sum_{r=1}^{t}\sum_{i,j\in T_r}(\alpha_i+\beta_i-C)(\alpha_j+\beta_j-C)\phi_r(\boldsymbol{x}_i)\phi_r(\boldsymbol{x}_j)$$
$$\text{s.t. }\sum_{i=1}^{N}\alpha_i y_i=0$$
$$\sum_{i\in T_r}(\alpha_i+\beta_i)=T_r C,r=1,...,t \tag{3}$$
$$\alpha_i\ge0,\beta_i\ge0,i=1,...,N$$

通过对上式的求解，可得 SVM+的决策函数：

$$f(\boldsymbol{x})=\boldsymbol{w}\cdot\phi(\boldsymbol{x})+b \tag{4}$$

2　多任务学习的不平衡 SVM+算法(MTL-IC-SVM+)

2.1　目标函数构造

由式（2）容易看到，SVM+算法在目标函数中追求训练样本错分的最小化，即

$$\min\sum_{i=1}^{N}y_i(\boldsymbol{w}\cdot\phi(\boldsymbol{x}_i)+b)-1 \tag{5}$$

在处理分类问题时，如果两类样本容量相差较大，分类面往往向少数类样本偏移来达到整体样本低错分率的目的。本文采取两类样本平均错分率最小原则，同时，为纠正分类面的偏移，寻找的分类面在达到两类间距离的最大化的同时保证少数类到分类面的距离不得小于多数类到分类面的距离，因此式(5)可改写成

$$\min\frac{1}{v^+N^+}(\sum_{i=1}^{N^+}y_i(\boldsymbol{w}\cdot\phi(\boldsymbol{x}_i)+b)-1)$$
$$+\frac{1}{v^-N^-}(\sum_{i=N^++1}^{N}y_i(\boldsymbol{w}\cdot\phi(\boldsymbol{x}_i)+b)-1-\rho^2) \tag{6}$$

其中:N^+ 和 N^- 分别是少数类和多数类样本的个数；v^+ 和 v^- 为两个正常数，用来调节两类样本的错分比例；常数 ρ^2 保证少数类到分类面的距离大于多数类到分类面的距离。

多任务学习的特性指多个任务中的数据一般属于多个分布不同但存在共性的数据域[10]，本文将 SVM+中的每个数据分组看成是一个子任务，自然地可以将该 SVM+改造成一个多任务学习模型。依据多任务学习方法的思想，多个子任务的决策模型应该是相似的，在保持各个子学习机局部优化的同时各学习机之间的全局差异最小化。此时，每个子任务的决策函数 f_r 可以表现为一个公共决策函数 g_0 和修正函数 g_r 的和：

$$fr=g0+gr \tag{7}$$

具体地，决策函数 f_r 可以写成

$$f_r(\boldsymbol{x})=\boldsymbol{w}\cdot\phi(\boldsymbol{x})+b+\boldsymbol{w}_r\cdot\phi_r(\boldsymbol{x})+d_r,r=1,...,t \tag{8}$$

其中:对于全体样本的决策函数 $g_0=\boldsymbol{w}\cdot\phi(\boldsymbol{x})+b$，对应于每个子任务的修正函数 $g_r=\boldsymbol{w}_r\cdot\phi_r(\boldsymbol{x})+d_r$ 。

基于以上的分析，给出 MTL-IC-SVM+算法的目标函数为

$$\min_{\substack{\boldsymbol{w},\boldsymbol{w}_1,...,\boldsymbol{w}_r,\rho_1,\\ ...\rho_r,b,d_1,...d_r}}\frac{1}{2}(\boldsymbol{w}\cdot\boldsymbol{w})+\frac{\gamma}{2}\sum_{r=1}^{t}(\boldsymbol{w}_r\cdot\boldsymbol{w}_r)+\sum_{r=1}^{t}(\frac{1}{v_r^+m_r^+}\sum_{i\in T_r^+}\xi_i^r)$$
$$+\sum_{r=1}^{t}(\frac{1}{v_r^-m_r^-}\sum_{j\in T_r^-}\xi_j^r)-v\sum_{r=1}^{t}\rho_r^2$$
$$\text{s.t. }\boldsymbol{w}\cdot\phi(\boldsymbol{x}_i)+b+\boldsymbol{w}_r\cdot\phi_r(\boldsymbol{x}_i)+d_r\ge1-\xi_i^r,$$
$$-(\boldsymbol{w}\cdot\phi(\boldsymbol{x}_i)+b+\boldsymbol{w}_r\cdot\phi_r(\boldsymbol{x}_i)+d_r)\ge1+\rho_r^2-\xi_j^r,$$
$$\xi_i^r\ge0,\xi_j^r\ge0,\quad i\in T_r^+,\ j\in T_r^-,r=1,...,t \tag{9}$$

其中：m_r^+ 和 m_r^- 分别表示少数类和多数类样本在第 r 个子任务中的样本个数，每个子任务中的数据规模不一。v_r^+ 和 v_r^- 分别对应第 r 个子任务中少数类和多数类的正则化常量。常数 γ 表示

决策函数和相关修正函数间的权重。ξ_i^r 和 ξ_j^r 分别表示少数类和多数类样本在第 r 个子任务中的松弛变量。

为了进一步阐述上述优化目标函数的机理，给出如下的分析与说明：

a)MTL-IC-SVM+算法在保证每个子任务学习达到最优的同时，需要考虑这 r 个子任务之间学习的相似性和一致性，以获取不同任务间有益的归纳信息。目标式中 $\sum_{r=1}^{t}(\boldsymbol{w}_r\cdot\boldsymbol{w}_r)$ 表示各个子任务之间的差异项，其数值越大，表示各任务之间的差异越大；反之，差异越小。惩罚的程度则用参数 γ 来调节。

b)公共决策函数和修正函数中使用的核函数可以相同，也可以不同。关于核函数的选择，本文在实验部分有详细介绍。

c)参照 SVM+对属性特征划分组的方式来产生子任务，MTL-IC-SVM+能够继承 SVM+利用样本的结构信息的特性，通过挖掘样本的隐藏信息来提高模型的泛化能力。

d)SVM+目标函数中，松弛变量表示为修正函数，由于松弛变量不得小于 0，所以修正函数也必须大于等于 0。而在 MTL-IC-SVM+算法中，修正函数表示为任务间的差异程度，因此修正函数无需设置为大于 0。

通过引入拉格朗日向量 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$,式（9）对应的拉格朗日函数可以写成以下形式：

$$\begin{aligned}&L(\boldsymbol{w},\boldsymbol{w}_1,\ldots,\boldsymbol{w}_t,\rho_1,\ldots,\rho_t,d_1,\ldots,d_t,\boldsymbol{\alpha},\boldsymbol{\beta})=\frac{1}{2}(\boldsymbol{w}\cdot\boldsymbol{w})\\&+\frac{\gamma}{2}\sum_{r=1}^{t}(\boldsymbol{w}_r\cdot\boldsymbol{w}_r)+\sum_{r=1}^{t}(\frac{1}{v_r^+m_r^+}\sum_{i\in T_r^+}\xi_i^r)+\sum_{r=1}^{t}(\frac{1}{v_r^-m_r^-}\sum_{j\in T_r^+}\xi_j^r)\\&-\sum_{r=1}^{t}(\sum_{i\in T_r^+}\alpha_i(\boldsymbol{w}\cdot\phi(\boldsymbol{x}_i)+b+\boldsymbol{w}_r\cdot\phi_r(\boldsymbol{x}_i)+d_r-1+\xi_i^r)\\&+\sum_{j\in T_r^-}\alpha_j(\boldsymbol{w}\cdot\phi(\boldsymbol{x}_j)+b+\boldsymbol{w}_r\cdot\phi_r(\boldsymbol{x}_j)+d_r+1+\rho_t^2-\xi_j^r))\\&-\sum_{r=1}^{t}\sum_{i\in T_r^+}\beta_i\xi_i^r-\sum_{r=1}^{t}\sum_{j\in T_r^-}\beta_j\xi_j^r-v\sum_{r=1}^{t}\rho_r^2\end{aligned}\tag{10}$$

根据 KKT 条件，可得

$$\frac{\partial L}{\partial\boldsymbol{w}}=0\Rightarrow\boldsymbol{w}=\sum_{i=1}^{N}\alpha_iy_i\phi(\boldsymbol{x}_i)\tag{11}$$

$$\frac{\partial L}{\partial\boldsymbol{w}_r}=0\Rightarrow\boldsymbol{w}_r=\frac{1}{r}\sum_{i=1}^{N}\alpha_iy_i\phi_r(\boldsymbol{x}_i)\tag{12}$$

$$\frac{\partial L}{\partial b}=0\Rightarrow\sum_{i\in T_r^+}\alpha_i-\sum_{j\in T_r^-}\alpha_j=0,r=1,\ldots,t\tag{13}$$

$$\frac{\partial L}{\partial\rho_r}=0\Rightarrow\sum_{j\in T_r^-}\alpha_j=v,\ \ r=1,\ldots,t\tag{14}$$

$$\frac{\partial L}{\partial\xi_i^r}=0\Rightarrow\alpha_i+\beta_i=\frac{1}{v_r^+m_r^+},i\in T_r^+,r=1,\ldots,t\tag{15}$$

$$\frac{\partial L}{\partial\xi_j^r}=0\Rightarrow\alpha_j+\beta_j=\frac{1}{v_r^-m_r^-},j\in T_r^-,r=1,\ldots,t\tag{16}$$

将式(11)~(16)代入(10),　　可得到式(10)的对偶式:

$$\begin{aligned}&\min_{\boldsymbol{\alpha}}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_jy_iy_j(\phi(\boldsymbol{x}_i)\cdot\phi(\boldsymbol{x}_j))\\&+\frac{1}{r}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_jy_iy_j(\phi_r(\boldsymbol{x}_i)\cdot\phi_r(\boldsymbol{x}_j))\\&\text{s.t.}\ \sum_{i\in T_r}\alpha_i\,y_i=0,r=1,\ldots,t\\&\sum_{i\in T_r^+}\alpha_i=v\,,\ \sum_{j\in T_r^-}\alpha_j=v\,,r=1,\ldots,t\\&0\le\alpha_i\le\frac{1}{v_r^+m_r^+},i\in T_r^+,r=1,\ldots,t\\&0\le\alpha_j\le\frac{1}{v_r^-m_r^-},j\in T_r^-,r=1,\ldots,t\\&\alpha_i\ge0,i=1,\ldots,N\end{aligned}\tag{17}$$

由式（17）易知，MTL-IC-SVM+对偶形式的时间复杂度为 $O(N^3)$，若采用 SMO 方法求解的时间复杂度为 $O(N^2)$。

2.2　v–性质分析

本节讨论 MTL-IC-SVM+模型中参数 v，v_1 和 v_2 参数之间的关系以及对训练精度的影响。根据 SVM 基本理论，一个训练样本 $\boldsymbol{x}_i(1\le i\le N)$ 如果其对应的松弛变量 $\xi_i>0$，那么这个样本称为错分样本。设 n_r^+ 和 n_r^- 分别表示第 r 个子任务中少数类和多数类错分样本的个数，s_r^+ 和 s_r^- 分别表示第 r 个子任务中少数类和多数类中支持向量的个数。

定理 1　vv_r^+ 和 vv_r^- 分别是少数类和多数类的错分率的上界和支持向量集的下界，即

$$\begin{aligned}&n_r^+/m_r^+\le vv_r^+\le s_r^+/m_r^+,\\&n_r^-/m_r^-\le vv_r^-\le s_r^-/m_r^-\end{aligned}\tag{18}$$

证明　式（17）中第 2 个约束项是 $\sum_{i\in T_r^+}\alpha_i=v$ ，根据 KKT 条件，所有 $\xi_i^r>0$ 的样本均满足 $\beta_i=0\ \ (i\in T_r^+,r=1,\ldots,t)$ 。从式（17）中第 3 个约束项可以看出，对每个任务中少数类样本中的每个错分样本均满足 $\alpha_i=1/v_r^+m_r^+\ \ (i\in T_r^+,r=1,\ldots,t)$ ，因此可得下式：

$$n_r^+/v_r^+m_r^+\le\sum_{i\in T_r^+}\alpha_i=v\tag{19}$$

此外，从式（17）可以看出，每个任务中的拉格朗日因子满足 $\alpha_i\le1/v_r^+m_r^+$ ，将这些 α_i 相加，可得：

$$\sum_{i\in T_r^+}\alpha_i\le s_r^+/v_r^+m_r^+\tag{20}$$

联合式（19）（20）可以得到不等式 $n_r^+/m_r^+\le vv_r^+\le s_r^+/m_r^+$ 。用类似的方法可以得证 $n_r^-/m_r^-\le vv_r^-\le s_r^-/m_r^-$ 。

3　实验与分析

依照不平衡分类问题中常用的设定方法，实验中将少数类指定为正类，将多数类指定为负类。为了评价 MTL-IC-SVM+的性能，实验将从两方面进行：a）针对决策函数和修正函数中核函数的选择的实验；b）与相关不平衡算法的比较性实验。实

验引入 SVM[19]、SVM+[15]、DEC[16]、EasyEnsemble[17] 和 AdaBoost[18]与本文所提算法进行了比较。这五种算法中，SVM 作为基线算法；DEC、EasyEnsemble 和 AdaBoost 均为不平衡分类算法，与之比较是为了验证本文算法与其他优秀的不平衡算法具有可比较甚至精度更高的性能。所有算法在 MATLAB2010b 环境下实现，SVM 算法由 LIBSVM 软件[19]实现。

3.1 实验设置

为体现不同程度的不平衡性对算法分类性能产生的影响，本文采用 G-mean 评价指标来评价算法的分类性能：

$$G-mean=\sqrt{Positive\ Accuracy\times Negative\ Accuracy}$$

其中: *Positive Accuracy* 为正类（少数类）样本的分类精度，*Negative Accuracy* 为负类（多数类）样本的分类精度。G-mean 指标因同时兼顾多数类和少数类样本的分类精度而被广泛用于处理不平衡数据分类问题。

参照文献[15,20]中的方法，实验中通过给属性划分数据组的方式来产生若干个多任务学习。鉴于医学数据集常出现类别的不平衡的现象，本节将在 4 个 UCI 医学数据集[21]上对 MTL-IC-SVM+进行评价。这四个 UCI 医学数据集分别是 Stalog Heart Disease（Heart），Pima Indians' diabetes（Pima），Hepatitis 和 BUPA Liver（Liver）。

Heart 集包含 13 个特征，实验中随机选择 40 个正类样本和 150 个负类样本构成 190 个样本的数据集，正负类比例是 4:15。首先，多任务学习 A 依据特征'age' 的分布范围将数据集划分成 3 个子任务：子任务 1 (age < 50，60 个样本)，子任务 2 (50≤age<60，66 个样本) 和子任务 3 (age≥60，64 个样本)。其次，Heart 集上多任务学习 B 依据特征'sex' 的分布范围将数据集划分成 2 个子任务：子任务 1 (sex = 0，47 个样本) 和子任务 2 (sex = 1，143 个样本)。

Pima 集包含 768 样本，8 个特征，其中正负类比例是 67:134。 Pima 集上多任务学习 A 依据特征 'age' 将数据集划分成 3 个不同的子任务：子任务 1 (age≤25，267 个样本)，子任务 2 (26≤age<39，294 个样本) 和子任务 3 (age≥40，207 个样本)。其次，多任务学习 B 依据特征'diabetes pedigree function'(pedigree)的分布范围划分 3 个子任务：子任务 1 (pedigree < 0.25，205 个样本)，子任务 2 (0.25≤pedigree≤0.5，286 个样本) 和子任务 3(pedigree >0.5，277 个样本)。

Hepatitis 集包含 19 个特征，实验中随机选择 30 个正类样本和 85 个负类样本构成 115 个样本的数据集，正负类比例是 6:17。实验中在这一数据集上产生两个多任务学习，多任务学习 A 依据特征'steroid' 的分布范围将数据集划分成 2 个子任务：子任务 1 (steroid=1，58 个样本)，子任务 2 (steroid=2，57 个样本)。多任务学习 B 依据特征 'malaise' 的分布范围将数据集划分成 2 个子任务：子任务 1 (malaise=1，61 个样本)和子任务 2 (malaise=2，54 个样本)。

Liver 数据集包含六个特征，其中正负类比例是 29:40，共 345 个样本。实验中多任务学习 A 的产生是依据特征'drinks number of half-pint equivalents of alcoholic beverages drunk per day'(drinks) 的分布范围将数据集划分成两个子任务：子任务 1 (drinks≤17，112 个样本)，子任务 2 (18≤drinks≤36，111 个样本) 和子任务 3 (drinks >36，112 个样本)。其次，Pima 数据集上多任务学习 B 是依据特征'sgpt alamine aminotransferase' (sgpt) 的分布范围将数据集划分成三个子任务：子任务 1 (sgpt≤20，104 个样本)，子任务 2 (21≤sgpt≤30，113 个样本) 和子任务 3 (sgpt > 31，118 个样本)。

本文使用 10 折交叉验证按照以下网格划分在训练集上寻找最优参数：SVM、SVM+和本文所提方法的高斯核的核参数在[0.1, 0.2, 0.4, 0.6, 1, 1.5, 3]，正则化参数 C 在[0.1, 1, 10, 100]，参数 γ 在[0.001, 0.1, 0.1, 1, 10]，参数 ν 在 [10, 30, 50, 70, 90]，参数 ν_r^+ 和 ν_r^- 在[0.001, 0.01]。对于其他对比算法，均按照原文参数设置方法完成设置，其中 DEC 中参数 C^-/C^+的值等于少数类样本容量与多数类样本容量的比值；对于 EasyEnsemble 和 Adaboost，设置弱分类器的个数是 10。

3.2 MTL-IC-SVM+中核类型的选择

正如前文所述，MTL-IC-SVM+算法中的决策函数和修正函数中的核函数是独立的，两者可以相同也可以不同。实验中分别在两者中使用线性核和高斯核 $\exp(-\sigma\|\boldsymbol{x}-\boldsymbol{y}\|^2)$，共有四种核类型组合，分别用符合 M1、M2、M3 和 M4 表示，如表 1 所示。高斯核的核参数 σ_1 和 σ_2 均在实验设定的范围内寻优获得。为了找到适用于 MTL-IC-SVM+的核类型，实验中分别将表 1 所示的四种核类型组合在 Heart、Pima、Hepatitis 和 Liver 集运行，结果如表 2 所示。

表 1 MTL-IC-SVM+中的核类型选择

核类型	决策函数中的核类型	修正函数中的核类型
M1	线性核	线性核
M2	线性核	高斯核 σ_2
M3	高斯核 σ_1	线性核
M4	高斯核 σ_1	高斯核 σ_2

表 2 MTL-IC-SVM+不同核类型下的 G-mean 值比较

数据集	多任务名称	M1	M2	M3	M4
Heart	多任务 A	70.13 ±1.01	72.86 ±1.00	76.62 ±1.14	**78.58** ±1.07
	多任务 B	71.00 ±1.12	72.59 ±1.06	76.08 ±1.20	**78.09** ±1.09
Pima	多任务 A	67.13 ±2.08	69.06 ±1.83	71.55 ±1.77	**73.09** ±2.18
	多任务 B	67.00 ±2.00	69.13 ±1.94	71.46 ±1.79	**73.12** ±2.02
Hepatitis	多任务 A	60.84 ±1.90	62.51 ±2.03	68.00 ±1.91	**68.86** ±1.66

Liver	多任务 B	60.32 ±1.78	62.11 ±2.21	67.89 ±1.85	**68.41** ±1.80
	多任务 A	61.06 ±2.56	61.99 ±2.12	65.77 ±2.30	**66.24** ±2.38
	多任务 B	61.15 ±2.74	61.84 ±2.33	65.82 ±2.85	**66.13** ±2.86

文献[3]得出结论：SVM 在绝大多数真实数据上使用非线性核的分类效果要优于使用线性核的情况。表 2 显 MTL-IC-SVM+在各数据集上的 G-mean 值最优值均在 M4 模型上获得，次优值在 M3 模型上获得，而在 M1 模型获得的 G-mean 值均是最低的。显然，本文所提 MTL-IC-SVM+在实验中验证了这一说法。因此在后续的实验中，本文在决策函数和修正函数上均使用高斯核函数。但需要说明的是，决策函数和修正函数中使用的高斯核函数使用不同的核参 σ_1 和 σ_2。

3.3 MTL-IC-SVM+性能比较

为了评价 MTL-IC-SVM+在不平衡分类问题中的性能，实验中将 MTL-IC-SVM+与 SVM、SVM+、DEC、EasyEnsemble 和 Adaboost 在四个不平衡 UCI 数据集上的性能进行了比较，实验结果如表 3 所示。从表中数据可以看出：

a）MTL-IC-SVM+对比 5 种对比算法在四个不平衡数据集上均取得了最好的 *G-mean* 值。实验中在每个数据集上均建立了两个任务学习任务，结果显示两者间的差距不大，说明不同的属性特征中均蕴含一定的样本结构信息。

b）较 SVM 和 DEC 只能将样本映射至决策空间，MTL-IC-SVM+可以将样本同时映射至决策空间和修正空间，这为 MTL-IC-SVM+适用于不同任务的训练数据提供了更多的灵活性。

c）SVM 和 SVM+未考虑数据的不平衡性造成分类面的偏移，从表中数据可知，这四种算法对应的 Positive accuracy 值较低，因此 *G-mean* 值也低于其他算法。

d）EasyEnsemble 和 Adaboost 算法使用过采样技术来增加少数类样本的数量，由于改变了样本的分布结构容易造成分类器过拟合的现象。因此，这两种算法获得的 *G-mean* 值也低于 MTL-IC-SVM+。

为了进一步评价 MTL-IC-SVM+在不同正负类比例下的分类性能，对四个 UCI 医学数据集进行改造，各类数据集随机划分成训练集和测试集，训练集包含从多数类样本中抽取的 70%样本和根据{20%，40%，60%，80%}不同取值所分别抽取的不同的少数类样本，其余样本作为测试数据集。考虑到 MTL-IC-SVM+中两个多任务分类效果相当，实验中在每个数据集上按照 4.1 节的组别的设置生成多任务 A，SVM+中的分组属性同样使用多任务 A 的分组属性。实验中依然通过 10 折交叉验证的方法进行参数的选择，图 1 记录了六种算法在四个不平衡 UCI 医学数据集上不同正负类比例下 *G-mean* 值。结果显示 MTL-IC-SVM+对于各数据集下不同的正负类比例均具有优良的分类性能。

表 3　UCI 数据集上不同分类器分类效果的比较

数据集	算法	Positive accuracy	Negative accuracy	G-mean
Heart	SVM	45.08±2.53	88.80±2.62	63.71±2.58
	SVM+ (分组属性‘age’)	47.19±2.05	88.00±2.42	64.47±2.16
	SVM+ (分组属性‘sex’)	47.10±2.47	87.75±1.74	64.29±2.02
	DEC	71.12±2.39	82.50±1.90	76.41±2.25
	EasyEnsemble	72.74±2.21	81.02±1.58	77.34±1.77
	Adaboost	72.75±2.42	81.51±1.55	77.22±1.88
	MTL-IC-SVM+ (多任务 A)	77.24±1.18	80.01±0.94	**78.58**±1.07
	MTL-IC-SVM+ (多任务 B)	76.19±1.08	80.44±0.88	**78.09**±1.09
Pima	SVM	50.10±2.34	88.30±2.45	66.57±2.35
	SVM+ (分组属性‘age’)	52.33±2.37	87.91±2.34	67.45±2.31
	SVM+ (分组属性‘pedigree’)	52.54±2.32	87.61±2.28	67.20±2.31
	DEC	65.94±2.83	77.93±2.34	71.29±2.52
	EasyEnsemble	68.12±2.56	76.80±2.11	72.64±2.46
	Adaboost	67.19±3.35	77.80±3.43	72.22±3.78
	MTL-IC-SVM+ (多任务 A)	71.05±2.13	75.80±2.06	**73.09**±2.18
	MTL-IC-SVM+ (多任务 B)	71.42±2.21	75.82±2.08	**73.12**±2.02

数据集	算法	Positive accuracy	Negative accuracy	G-mean
Hepatitis	SVM	35.76±1.07	97.75±2.24	59.03±1.64
	SVM+ (分组属性‘steroid’)	39.43±2.47	95.58±1.96	61.16±2.07
	SVM+ (分组属性‘malaise’)	39.86±2.53	95.27±2.31	61.26±2.44
	DEC	56.63±1.52	88.02±2.60	68.08±2.21
	EasyEnsemble	55.45±1.01	83.35±1.60	67.75±1.47
	Adaboost	55.35±2.72	83.41±1.60	67.70±2.01
	MTL-IC-SVM+ (多任务 A)	57.32±1.75	81.76±1.50	**68.86**±1.66
	MTL-IC-SVM+ (多任务 B)	57.15±1.24	81.76±2.13	**68.41**±1.80
Liver	SVM	40.32±2.98	75.34±2.45	54.20±2.68
	SVM+ (分组属性‘drinks’)	43.79±2.74	72.85±2.24	56.11±2.36
	SVM+ (分组属性‘sgpt’)	43.29±2.73	73.03±2.44	55.01±2.66
	DEC	60.87±2.14	71.03±2.34	65.35±2.22
	EasyEnsemble	60.03±2.32	71.56±2.62	65.37±2.47
	Adaboost	60.57±2.77	71.32±2.81	65.54±2.79
	MTL-IC-SVM+ (多任务 A)	62.41±3.69	70.84±2.53	**66.24**±2.38
	MTL-IC-SVM+ (多任务 B)	61.83±2.30	71.52±2.42	**66.13**±2.86

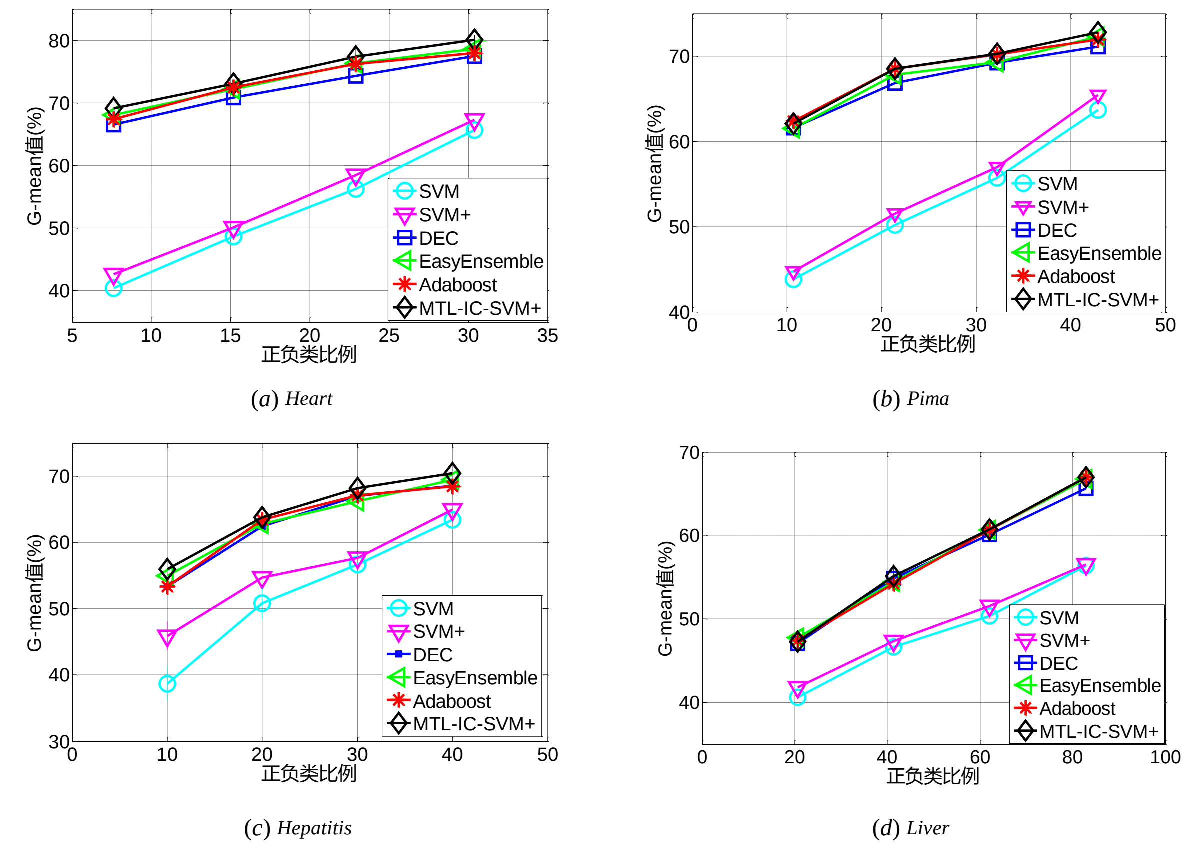


图 1　UCI 医学集上 6 种算法在不同正负类比例下 G-mean 比较

4 结束语

本文提出的 MTL-IC-SVM+在使用“大间隔”的机制设置少数类到分类面的距离大于多数类到分类面的距离，并按照样本数比例设置多数类和少数类样本不同的错分惩罚因子的同时，将 SVM+的分组挖掘样本隐藏信息的单任务学习改造为多任务学习的模型来提高模型的分类泛化能力。在 4 个不平衡 UCI 数据集上的实验表明，MTL-IC-SVM+具有良好的分类性能。应当指出，本文对如何更合理地选择特征属性作为划分子任务的依据，以及 MTL-IC-SVM+能否有效解决大样本、处理有噪声数据等问题没有进行深入探讨，MTL-IC-SVM+仍面临进一步提高实用性的挑战，这些将作为笔者近期的研究重点。

参考文献：

[1] Sun Zhongbin, Song Qinbao, Zhu Xiaoyan, *et al.* A novel ensemble method for classifying imbalanced data [J]. Pattern Recognition, 2015, 48 (5): 1623-1637.

[2] 刘东启，陈志坚，徐银，等. 面向不平衡数据分类的复合 SVM 算法研究 [J]. 计算机应用研究，2018, 35 (4): 1023-1027. (Liu Dongqi, Chen Zhijian, Xu Yin, *et al.* Hybrid SVM algorithm oriented to classifying imbalanced datasets, Application Research of Computers, 2018, 35 (4): 1023-1027.)

[3] Chen Wei, Pourghasemi H, Kornejady A, *et al.* Landslide spatial modeling: introducing new ensembles of ANN, MaxEnt, and SVM machine learning techniques [J]. Geoderma, 2017, 305 (11): 314-327.

[4] Wang Zhigang, Zhao Zengshun, Weng Shifeng, *et al.* Incremental multiple instance outlier detection [J]. Neural Computing & Applications, 2015, 26 (4): 957-968.

[5] Zhao Zengshun, Feng Xiang, Wei Fang, *et al.* Learning representative features for robot topological localization [J]. International Journal of Advanced Robotic Systems, 2013, 10 (4): 1-12.

[6] Abidine M, Fergani B. Effect of oversampling versus undersampling for SVM and LDA classifiers for activity recognition [J]. International Journal of Design & Nature & Ecodynamics, 2016, 11 (3): 306-316.

[7] Maldonadoa S, López J. Dealing with high-dimensional class-imbalanced datasets: Embedded feature selection for SVM classification [J]. Applied Soft Computing, 2018, 67 (6): 94-105.

[8] Wang Boyu, Pineau J. Online bagging and boosting for imbalanced data streams [J]. IEEE Trans on Knowledge and Data Engineering, 2016, 28 (12): 3353-3366.

[9] López V, Fernández A, Jesus M, *et al.* A hierarchical genetic fuzzy system based on genetic programming for addressing classification with highly imbalanced and borderline datasets [J]. Knowledge Based Systems, 2013, 38 (3): 85-104.

[10] Yu Lean, Zhou Rongtian, Tang Ling, *et al.* A DBN-based resampling SVM ensemble learning paradigm for credit classification with imbalanced data Research article [J]. Applied Soft Computing, 2018, 69 (8): 192-202.

[11] Cheng Fanyong, Zhang Jing, Wen Cuihong, *et al.* Large Cost-Sensitive Margin Distribution Machine for Imbalanced Data Classification [J]. Neurocomputing, 2017, 224 (8): 45-57.

[12] Jiang Yingzhang, Deng Zhaohong, Chung F L, *et al.* Multi-task TSK fuzzy system modeling by mining inter-Task common hidden structure [J]. IEEE Trans on Cybernetics, 2015, 45 (3): 548-61.

[13] Jiang Yingzhang, Deng Zhaohong, Choi K S, *et al.* A novel multi-task TSK fuzzy classifier and its enhanced version for labeling-risk-aware multi-task classification [J]. Information Sciences, 2016, 357 (2): 39-60.

[14] Vapnik V, Vashist A. A new learning paradigm: Learning using privileged information [J]. Neural Networks, 2009, 22 (5): 544-557.

[15] Liang Lichen, Cai Feng, Cherkassky V. Predictive learning with structured (grouped) data [J]. Neural Networks, 2009, 22 (6): 766-773.

[16] He Haibo, Ma Yunqian. Imbalanced learning: foundations, algorithms, and applications [M]. Hoboken: Wiley, 2013: 83-96.

[17] Liu Xuying, Wu Jianxin, Zhou Zhihua. Exploratory undersampling for class imbalance learning [J]. IEEE Trans on Systems, Man, and Cybernetics, Part B: Cybernetics, 2009, 39 (2): 539-550.

[18] Wang Shuo, Yao Xin. Multiclass imbalance problems: analysis and potential solutions [J]. IEEE Trans on Systems, Man and Cybernetics Part B: Cybernetics, 2012, 42 (4): 1119-1130.

[19] Chang C C, Lin C J, LIBSVM: a library for support vector machines [J]. ACM Trans on Intelligence System Technology, 2011, 2 (3): 1-27.

[20] Zhu Wenxin, Zhong Ping. A new one-class SVM based on hidden information [J]. Knowledge Based Systems, 2014, 60 (4): 35-43.

[21] UC Irvine Machine Learning Repository. UCI database [DB/OL]. [2016-09-28] http://www. ics. uci. edu/%20mlearn/MLRepositoy. html.